DOCUMENT RESUME

ED 441 862 TM 030 898

AUTHOR Kalohn, John C.; Spray, Judith A.
TITLE Test Security and Item Exposure Control for Computer-Based Examinations: Performance of a Computerized Classification Test for Professional Certification.
PUB DATE 2000-04-00
NOTE 11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Certification; *Classification; *Computer Assisted Testing; *Licensing Examinations (Professions); Test Items; Test Use
IDENTIFIERS *Item Exposure (Tests); Paper and Pencil Tests; Sequential Probability Ratio Test (Wald); *Test Security

ABSTRACT

A client of American College Testing, Inc. (ACT) decided to implement a computer-based testing program to replace their paper-pencil format for professional certification. This paper reports on the results of the developed test after 1 year's use, especially as the results relate to test security issues. ACT research shows that a variable length computerized classification test using the sequential probability ratio test worked well in retaining the estimated classification accuracy of the paper-pencil test. The developed examination consisted of a minimum of 100 scored items with a maximum of 120 scored items, with 25 pretest items administered throughout the examination. Scale scores were obtained on the total test and two subtest content domains using maximum likelihood estimates of theta after testing had ended. As of February 2000, 1,751 candidates had taken the examination, and of this number, 1,524 (87%) had passed the examination. These initial results, with about half of the expected candidate population tested, suggest that the item exposure control parameters that were calculated during the initial simulations worked appropriately. This conclusion is based on the assumption that the previous paper-pencil test-taking behavior will continue with the computerized version in that higher ability test-takers will take the examination in the fall and the less able will sit for the examination in the winter or spring. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 862

J Kalohn

2A

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Test Security and Item Exposure Control for Computer-based Examinations: Performance of a Computerized Classification Test for Professional Certification

John C. Kalohn and Judith A. Spray

ACT, Inc.

TM030898

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. April 2000

BEST COPY AVAILABLE

Test Security and Item Exposure Control for Computer-based Examinations: Performance of a Computerized Classification Test for Professional Certification

Six years ago, an ACT client decided to implement a computer-based testing program to replace their paper/pencil format for professional certification. Extensive research of the their item pool was conducted to determine if a CBT program could replace the traditional paper/pencil test form and, if so, what kind of CBT would be optimal, given the current status of the item pool and past performance of the paper/pencil test. Based on the results of this research, the client initiated a CBT program for their exam this past year. The purpose of this paper is to report on the results of the testing program thus far, particularly as these results relate to test security issues. In particular, this paper answers the following question: Do the item and test security measures initiated appear to be working as the program enters its final months of testing for the first annual CBT cycle?

Background

The paper/pencil exam had been a fixed-form, fixed-length test containing 240 scored test items and 20 unscored pretest items. The test was given twice a year. The test volume in fall administration (1) was always larger than the spring administration and (2) consisted of candidates who were more able than those who sat for the exam in spring. For example, in 1998, the last year in which the paper/pencil exam was given, there were 2640 candidates who sat for the exam in the fall compared to only 1139 for the spring administration. Of those who took the exam in the fall, 83% passed the exam, compared to the spring passing rate of only 68%. Overall, the passing rate for 1998 was 78%.

The challenge in making the transition from paper/pencil to a CBT exam was to retain the estimated classification accuracy that existed in the current version of the paper/pencil test, regardless of the type of CBT program recommended. Thus, one of the first tasks in

terms of researching the viability of CBT for this program was to determine if a CBT exam could obtain the estimated classification accuracy produced by the fixed-length paper/pencil exam and, if so, what type of CBT was optimal. Because the exam was to be a computerized classification test (CCT), ACT compared two CBT classification formats: (1) a CCT using the sequential probability ratio test or SPRT (Wald, 1947; Ferguson, 1969); and (2) a CCT based on a traditional CAT algorithm in which examinee ability (θ) was estimated throughout testing, items were administered at the updated estimates of θ, and a final estimate of θ was compared to the latent passing score, θ_p. ACT's research showed that a variable-length CCT using the SPRT approach worked better than the traditional CAT algorithm in retaining the estimated classification accuracy from the paper/pencil test, and ultimately this approach was recommended to the client as optimal for their exam.

CCT Exam Characteristics

The variable-length CCT exam consists of a minimum of 100 scored items with a maximum of 120 scored items. Twenty-five pretest items are also administered throughout the examination. Candidates are not allowed to skip items and cannot return to previously answered questions. Scale scores are obtained on the total test and two subtest content domains using maximum likelihood estimates (MLEs) of θ after testing has ended.

The SPRT procedure used at ACT for licensure and certification exams selects items on the basis of: (1) item information at θ_p (Kalohn & Spray, 1999; Spray & Reckase, 1996; Spray, 1993); and (2) content requirements dictated by the content outline or test blueprint. Because the SPRT CCT for this exam is variable in length, it is not possible to meet the target content specifications exactly. A method for item selection implemented by Davey & Fan (2000) was used in conjunction with a

Sympson-Hetter (1985) approach to meet content specifications within a given range and to determine item exposure parameters for the 777 items in the item pool, or P(A|S), the probability that an item will be administered, given that it is selected for administration.

This item exposure control algorithm uses a penalty function to prioritize the selection of items for administration in a CCT based on the examination's content specifications and each item's information at the latent passing score. During the item selection process the algorithm determines what items in the pool are available for administration to the examinee. Each item receives a weight that is based upon the item's content classification. The weighting of the content is variable based upon where in the testing process that examinee is located and what the relative demands of the content specification are at that point in the examination. In other words, as items are administered, the selection of the next item is affected by how the content specifications are currently being met by the test. This is critically important for variable length tests and assists in approaching the content specifications throughout the administration of the examination. In addition to content specifications, item information at the passing score level is also considered and impacts which item is selected next so that informative items receive a higher priority than low information items.

Item exposure controls were set using a modification of the Sympson-Hetter (1985) approach. Since the CCT is the simplest case of a CAT in the general case, items are selected based on the information of the item at the latent passing score, and concerns associated with preventing high overlap of items at various points along the θ metric are of no concern. The primary modification of the Sympson-Hetter approach is that starting exposure control parameters are set at the maximum target rather than 1.0. Our research has shown that we can reduce the number iterations taken to reach stable item exposure control parameters. Using this process reduces the number of iterations from 25-30 to 6-8 with the initial exposure control parameter set to 1.0 and .20 (the maximum target for this examination), respectively.

Simulation Results

Final simulations on the item pool resulted in the estimation of item exposure parameters for each item in the pool. A target exposure rate of .20 was used to establish these parameters. Thus, no single item was to be exposed or administered more than 20% of the time. Because the candidate population consisted of two distinctly different distributions in terms of their ability, the simulations had to employ a mixed ability distribution, $g(\theta)$. Thus, $g(\theta)$ was a mixed normal distribution, estimated from previous paper/pencil forms administered in fall and spring, with

$$g(\theta) = \alpha g_1(\theta) + (1-\alpha) g_2(\theta), \quad (1)$$

where α was a proportional mixing constant, and $g_1(\theta)$ and $g_2(\theta)$ were normal distributions representing the fall and spring candidate populations, respectively and α representing the mixing proportion of the two candidate populations.

The final simulations were run for 5000 examinees with $\theta \sim g(\theta)$. A sample of the simulation output can be seen in Figure 1. In evaluating the validity of the simulation process, ACT compares these predicted results to those observed from the live CCT at the end of a full testing year.

Results as of February, 2000

As of February of this year, 1751 candidates had taken the exam. Of this number, 1524 or 87% had passed the exam. The average test length of these candidates was 106 items with a standard deviation of approximately 9 items. When compared to the simulation output in Figure 1, it is important to recognize that only about half of the expected candidate population has taken this exam. Because the CBT form of the exam

began in the summer of 1999, there is a good chance that those candidates who have taken the exam tended to be from the *more able* rather than the *less able* population.

Simulation of SPRT CCT

DATE: 2/02/1999 **TIME:** 13:30

Test Plan Control File:	control
Parameter Filename:	122398.ip
Item Information Filename:	122998.inf
Exposure Control Filename:	18.exp
Output Filename:	r18.0ut
Enemy Count Filename:	Encount.dat
Enemy List Filename:	Enfile.dat
Theta Distribution Filename:	sample.1
Passing Theta:	-0.69
Upper Bound:	-0.49
Lower Bound:	-0.89
Power:	0.95

Test Specifications and Results

Domain	Target	Mean(SD)
1	0.15	0.150(.001)
2	0.40	0.392(.003)
3	0.05	0.059(.002)
4	0.20	0.199(.002)
5	0.20	0.200(.001)

Simulation Results

# of simulated examinees:	5000
Passing Rate:	0.694
Failing Rate:	0.306
False Positive Error Rate:	0.038
False Negative Error Rate:	0.041
Total Error Rate:	0.079
Average Test Length:	107.436
SD Test Length:	9.467
Minimum Test Length:	100
Maximum Test Length:	120

Figure 1. Sample SPRT CCT Simulation Results

Of course, only time will tell if this is a true assumption. If it *is* true, then we would expect the passing rate to drop and the average test length to increase slightly. The latter expectation is based on current data that show that failing candidates tend to be classified

BEST COPY AVAILABLE

after taking the maximum number of test items (i.e., 120) more often than any other number of test items. Thus, this assumption if true, should tend to increase the average test length and increase the variability of test length as well.

We are also interested in monitoring (1) actual item exposure rates, (2) actual numbers of exposures, and (3) observed p-values for each item. The first and third values can be compared to the expected or predicted values to determine whether the item selection algorithm and the predicted values of P(A|S) are operating properly. The actual number of item exposures will be compared to some upper limit to determine when an item should be retired or temporarily set aside.

The actual item exposure rates can be seen in Figure 2. The vertical line marks the upper limit set during the simulation runs, or .20. Of the 777 items, only 5 had actual observed exposure rates greater than .20. For the 640 items that have been used in this exam the mean observed exposure rate was .165 with a standard deviation of .04. The maximum observed item exposure rate was .207 for an item with 363 administrations. It should be noted that there are 141 items with 0 or 1 exposures. These are items that provide little, if any, information at the passing score level and are never (or rarely) selected for administration. The maximum number of item exposures thus far is 363.

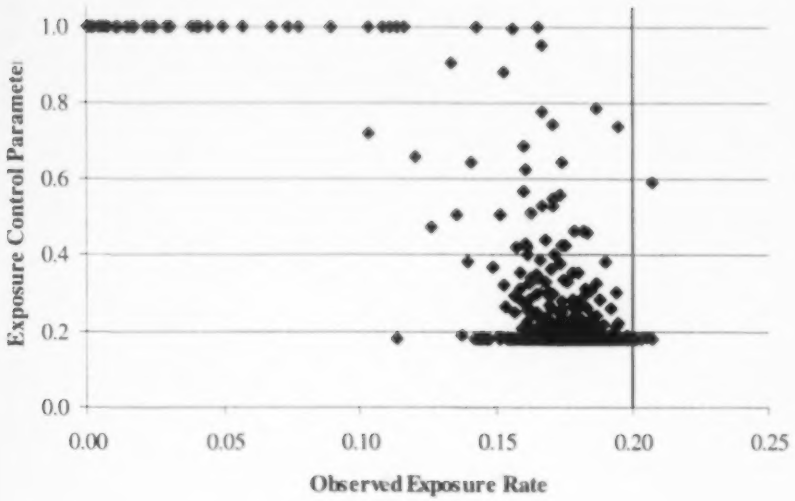

Figure 2. Observed Item Exposure

In terms of monitoring the performance of the items, the estimated p-values are compared to the observed p-values as an added check of item security. This comparison is more meaningful when all of the test data for the year have been collected and analyzed. If, as expected, only the most able examinees have tested thus far, observed p-values would be significantly higher than the predicted values. The predicted or expected p-values, $E\{P(\theta)\}$, were determined from

$$E\{P(\theta)\} = \int_{-\infty}^{\infty} P(\theta) g(\theta)\, d\theta\,, \tag{2}$$

where, $P(\theta)$ was the estimated item characteristic function from a 3-parameter logistic function, and $g(\theta)$ was the mixed normal distribution described previously.

For the purposes of calculating the expected p-values from equation (2) above, the two normal distributions, $g_1(\theta)$ and $g_2(\theta)$, were assumed to be N(0,1) and N(-.5,.81) with $\alpha = .70$, which were values obtained from a previous paper/pencil administration. The observed p-values were then compared to the predicted p-values as observed in Figure 3. Once again, it is expected that the observed p-values should decrease as the testing cycle ends in the summer of 2000.

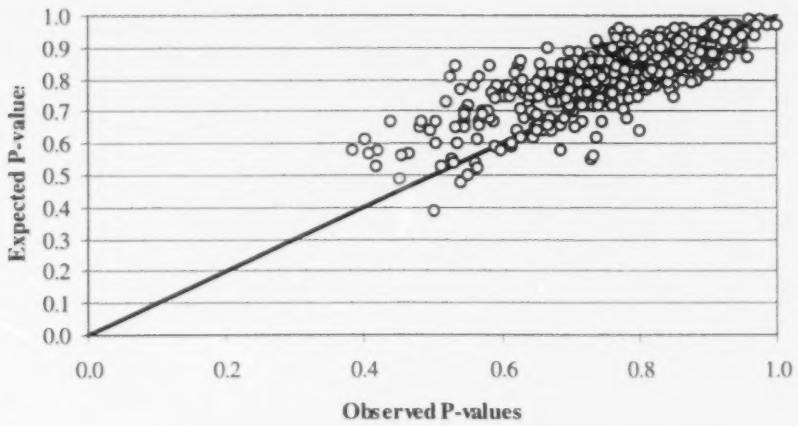

Figure 3. Observed P-values versus Predicted P-values

Conclusions

Based on these initial results with about half of the expected candidate population tested, it would appear that the item exposure control parameters that were calculated during the initial simulations are working appropriately. This conclusion is based on the assumption that the previous paper/pencil test-taking behavior observed will continue with the CBT version, in that the higher ability test-takers will take the exam in the fall while the less-able test-takers sit for the exam in the winter or spring months. If this assumption does not appear to hold, new simulations may have to be conducted based on a new distribution of population of test-takers (e.g., the higher ability test-takers with $\theta \sim N(0,1)$) .

References

Ferguson, R. L. (1969). *Computer-assisted criterion-referenced measurement* (Working Paper No. 41). Pittsburgh, PA: University of Pittsburgh, Learning and Research Development Center. (Eric Document Reproduction Series No. ED 037 089).

Davey, T., & Fan, M. (2000). *Specific information item selection for adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Kalohn, J. C., & Spray, J. A. (1999). The effect of model misspecification on classification decisions made using a computerized test. *Journal of Educational Measurement,* 36, 46-58.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21,* 405-414.

Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (ACT Research Rep. No. 93-7). Iowa City, IA: ACT.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Wald, A. (1947). *Sequential Analysis.* New York: Wiley.